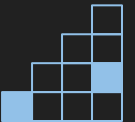




Data augmentation for NLP

Min/Jayadev

29 May, 2019



What is Data Augmentation?

- Technique to increase the amount of relevant data
- More data usually means better accuracy
- Very useful when you have small training datasets
- Data Augmentation is popular in computer vision
 - Images are shifted, zoomed in/out, rotated, flipped, distorted, or shaded with a hue [1]
- **What about natural language data?**

[1] Perez, Luis & Wang, Jason. (2017). The Effectiveness of Data Augmentation in Image Classification using Deep Learning.

Training data vs. accuracy

Sentiment Classification of IMDB movie reviews (Logistic Regression)

- Model performance gets better with increase in training data

Number of Training Examples	Model Test Accuracy
1000	0.79432
2000	0.82632
25000 (All)	0.86592

- Small num. of training examples => poor performance (generally)
- Data augmentation can help boost performance!

How do we augment natural language training data?

We will discuss 2 approaches today:

- Easy Data Augmentation
 - Use simple heuristics to augment training data
- Back translation
 - Use noise introduced by NMT to augment training data

Baseline: IMDB 1000 examples trained Model
(Acc: **0.79432**)

- Let's augment the data using the above approaches

Final projects

The final project is the main assignment of the course. Projects are required to be related in a substantive way to at least one of the central topics of the course. Final projects can be done in groups of 1–3 people; in our experience, groups of 3 lead to the best outcomes, so we encourage you to form a team of that size.

Each project team will be assigned a mentor (a member of the teaching team), who will provide feedback on all their project-related work and generally be available as a resource.

The final project is the main assignment of the course.

We will use this sentence to explore data augmentation methods!

Easy Data Augmentation [Wei and Zou, 2019]

<https://arxiv.org/abs/1901.11196>

4 simple techniques for data augmentation in NLP

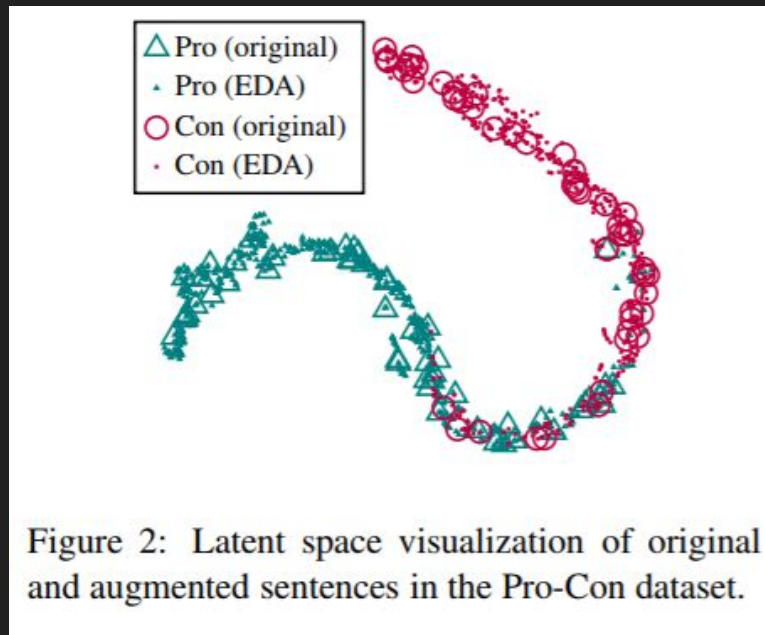
- **Synonym Replacement (SR):**
 - Randomly choose n words from the sentence that aren't stop words, replace with synonyms
- **Random Insertion (RI):**
 - Find a random synonym of a word that is not a stop word, and insert this randomly n times
- **Random Swap (RS):**
 - Swap two random words in the sentence, do this n times
- **Random Deletion (RD):**
 - Randomly remove each word in the sentence with probability p

Examples ($n = 1$)

- Original sentence: The final project is the main assignment of the course.
- SR: The final project is the **principal** assignment of the course.
- RI: The final project is the main assignment of **last** the course. (last~final)
- RS: The final project is the **assignment main** of the course.
- RD: The final is the main assignment of the course. (**project** deleted)

Do all these sentences preserve meaning? Not necessarily the case!

Do sentences retain “meaning” after EDA?



RNN representations of modified sentences are very similar those of original ones.

[Wei and Zou, 2019] (<https://arxiv.org/pdf/1901.11196.pdf>)

EDA Experiment Results

EDA (https://github.com/jasonwei20/eda_nlp) performs:

- Synonym Replacement
- Random Insertion
- Random Deletion
- Random Swap

Model Training Data	Accuracy
1000 training examples	0.79432
1000 training examples + 10000 EDA augmented examples	0.80348




- **~ 1.2% accuracy improvement on IMDB movie review sentiment classification**

Back translation (Sennrich et al., 2016)



- Using Neural Machine Translation to augment training data
 - Original Language -> Intermediate Language -> Original Language
 - Introduce noise through translation to get similar sentences

<https://arxiv.org/pdf/1511.06709.pdf>

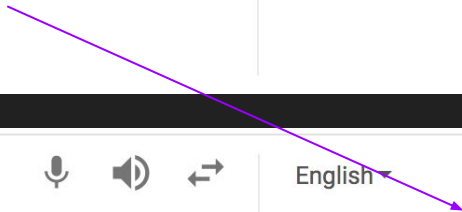
Example 1: English -> German -> English




English ▾   

The final project is the main assignment of the course. [Edit](#)



German ▾  

Das Abschlussprojekt ist die Hauptaufgabe des Kurses.



German - detected ▾   

Das Abschlussprojekt ist die Hauptaufgabe des Kurses. [Edit](#)

English ▾  

The final project is the main task of the course.

Example 2: English -> Korean -> English

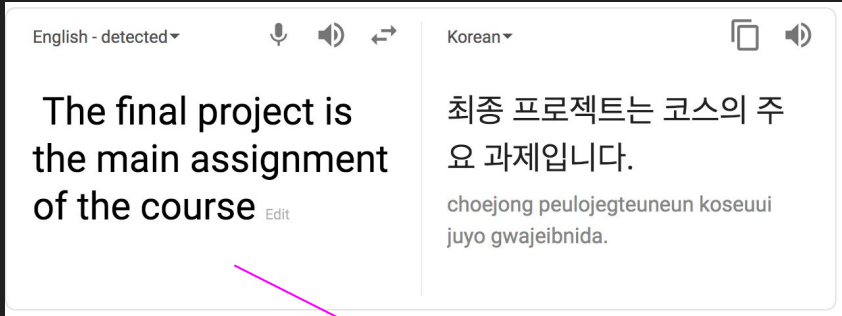
English - detected

The final project is the main assignment of the course Edit

Korean

최종 프로젝트는 코스의 주요 과제입니다.

choejong peulojegteuneun koseui juyo gwajeibnida.



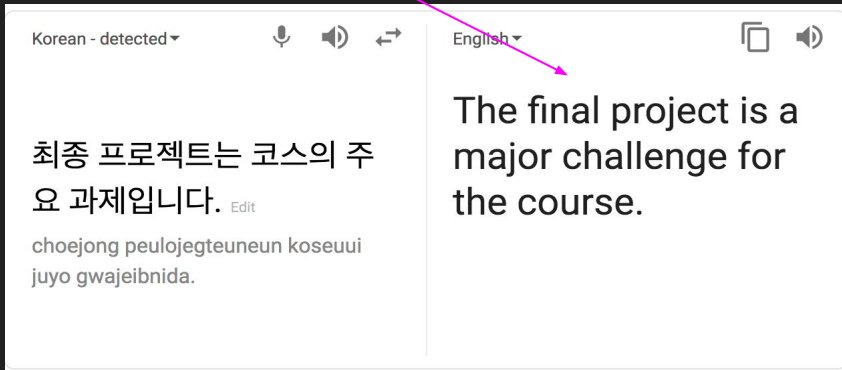
Korean - detected

최종 프로젝트는 코스의 주요 과제입니다. Edit

choejong peulojegteuneun koseui juyo gwajeibnida.

English

The final project is a major challenge for the course.



Back Translation Experiment Results

Model Training Data	Accuracy
1000 training examples	0.79432
1000 training examples + 1000 back-translated examples	0.80856

- ~ 1.8% accuracy improvement on IMDB movie review sentiment classification
- Used Google Translate API to translate from English -> German -> English

Back Translation + EDA

When trained on 12000 training examples:

- **1000** original training examples
- **1000** augmented from back translation
- **10000** augmented from EDA

Accuracy: **0.81092**

- ~ 2% accuracy increase